EDITORIAL

E-science and biomedical libraries

DOI: 10.3163/1536-5050.97.3.001

For a few years now, segments of the research library community have been carefully tracking the emergence of e-science and exploring its implications for the future of research libraries. During 2007–2008, I had the opportunity to work with the Association of Research Libraries (ARL) and a task force looking at issues in e-science and library support for research in the sciences. I also had the opportunity to reflect on how these phenomena are the same or different in the health sciences arena. I will explore some of those ideas here. First, what is e-science, how is it likely to affect libraries, and how are libraries positioned to respond to it? What are the major e-science initiatives in the sciences and engineering and in biomedicine? And finally, how does it translate to the health sciences library context?

What is e-science?

E-science has been described as a new research methodology, fueled by networked capabilities and the practical possibility of gathering and storing vast amounts of data. E-science can be distinguished from well-established experimental and theoretical methodologies by its large-scale, data-driven, and computationally intensive characteristics [1]. E-science alters the types of problems that scientists address, the tools that they use, and the nature of the publication that results from their research. Instead of conducting research to collect and analyze data, a typical e-science scenario mines existing data in search of patterns or correlations. Findings may support or undermine a hypothesis or lead to other questions and more mining. Many disciplines that used to be data poor now have more data than they know what to do with, thanks to sensor networks, advanced instrumentation, and increased computing power and storage capacity.

How is e-science likely to affect libraries?

The ARL e-science task force report [2] stated that e-science requires new strategies for research support and significant development of library infrastructure. Nearly all aspects of the research library's functions and roles may need to change to support these new methodologies. E-science tends toward inter- and multidisciplinary approaches that depend on computation and computer science. Research libraries have traditionally been discipline focused and, although increasingly technologically sophisticated, do not have systems of the scale or complexity of the e-science environment. E-science is data intensive, but research libraries have not typically been responsible for scientific data. E-science is frequently conducted in a team context, often distributed across multiple institutions and on a global scale. The primary constituency of libraries generally comprises those affiliated with the local institution. Licenses for electronic content are typically restricted to a particular institutional community, and the infrastructure to move institutional licenses into a multi-institutional environment is not well developed. E-science challenges all these traditional paradigms of research library organization and services.

Other areas have an easier fit. Research libraries already have existing capacity and expertise that they can bring to bear to support e-science. For example, research libraries have (1) expertise in the policies and principles related to open access and exchange of research information, as well as in the roles that can be played by institutional repositories to assure that exchange; (2) expertise in developing and supporting integration and interoperability tools (e.g., link resolvers, metasearch, metadata standards); (3) experience developing and supporting both business and technical strategies for long-term archiving (e.g., archival support generally, Portico); and (4) understanding of the archival and life-cycle aspects of scientific information, including the importance of assuring access and usability over the long term (preservation, metadata) [2].

What are the major e-science initiatives?

In the science and engineering communities, the National Science Foundation (NSF) is the primary source of research funding and the primary sponsor of cyberinfrastructure and e-science initiatives. NSF's vision for an infrastructure to support e-science contains four interdependent areas of investment: (1) high-performance computing; (2) data, data analysis, and visualization; (3) virtual organizations for distributed communities; and (4) learning and workforce development [3]. Research libraries have an interest in all these areas, and the last three offer particular opportunities: investments in data, metadata, ontologies, data collections, and development of a national digital data framework; investments in tools and technology systems for collaboration as well as evaluative research on the social and organizational dimensions of virtual communities; and investments to prepare professionals who will support, deploy, develop, and design cyberinfrastructure [2].

Potentially the single most significant driver of research library collaboration with e-science activities to date is the NSF's solicitation known as DataNet [4]. At the heart of the NSF DataNet vision is a new type of organization based on the research library as a model for stewardship, sustainability, and focus on user needs. This new type of organization is the research library transformed through a deep connection with, and understanding of, the needs and mores

of the scientific research community and through deep alliances with other entities involved in cyberinfrastructure-enabled research.

What is the role for biomedical libraries in e-science?

The biomedical and health sciences communities are familiar with the predominant role that the National Institutes of Health (NIH) play in funding research and fostering a biomedical research agenda. Somewhat analogous to the NSF DataNet development is the NIH Clinical and Translational Science Awards (CTSA) program that seeks to transform the conduct of clinical and translational research in order to yield new treatments more efficiently and quickly in part by designing new and improved clinical research informatics tools [5].

The role for health sciences libraries in the CTSA projects that have been funded to date is much less explicit than the role designated for research libraries in the DataNet initiative. For one thing, development of a sustainable data network is not a stated goal of CTSAs, so the concept of the library as a data steward is absent. However, several health sciences libraries are involved in the CTSA projects at their institutions. Anecdotal evidence indicates varying levels of collaboration occurring in designing, developing, and testing informatics tools for data management funded by CTSA. Several projects feature the development or refinement of ontologies and registries to enable interoperability among clinical data systems. The CTSA program at the University of Washington in Seattle, for example, includes projects on access to electronic health data, clinical data management, access to scientific instrumentation data, and clinical data integration. Librarians have opportunities here to collaborate with biomedical informatics and clinical researchers. Although we in the health sciences library community tend not to explicitly label this as e-science, it is directly analogous to e-science work in the science and engineering domains.

The health sciences library community is well aware of the role played by NIH, through the National Library of Medicine (NLM), in conducting the research and development and providing the infrastructure that the national health sciences library community relies on. But most of us are less aware that our colleagues in the science and engineering library communities have no such infrastructure or research and development effort to rely on. Without this lead, the national community of research libraries must attempt to fill the gap by forming a federation and relying on the strengths of individual institutions and consortia of institutions.

This lack of focused support at the federal or national level for science and engineering libraries (although the Library of Congress does provide some of this) is an obvious disadvantage in that institutions must then invest their own scarce resources in the research and development needed to create systems, tools, resources, standards, and policies to support new forms of digital scholarship. At the same time, it also presents an advantage: through their investment, these institutions and consortia develop expertise in supporting e-research. With many exceptions, of course, the health sciences library community does not tend to develop this expertise internally because it does not have to. Much of the community can rely on the tools and resources developed by NLM and use those tools and resources to support researchers and clinicians. One might describe the health sciences library community as being more focused on user needs and uses, because it can be; whereas the broader research library community is more consumed with developing the systems, tools, and resources they need to deliver e-science content and services, because they have to.

The powerful and definitive role of NLM in this regard is evident in the development of the Entrez cross-database search system. Entrez is a powerful federated search engine that allows users to search many discrete databases built by the National Center for Biotechnology Information (NCBI). Entrez is an early, and still leading, example of an advanced search and analysis tool that allows researchers to search across multiple databases, bringing together diverse information sources such as journal citations and abstracts, subject headings, the full text of journal articles and books, protein sequences, and gene sequence data. Entrez also provides an application programming interface (API) that enables a structured interface to all the databases and a flexible set of tools to discover unexpected patterns in the databases [6].

Many resources and tools developed by NCBI/NLM have become central elements in the progress of molecular biology and proteomics/genomics research. One interesting example—linking disparate sources with potentially far-reaching ramifications—is the database of Genotype and Phenotype (dbGaP), which is designed to explore the association between specific genes and observable traits or the presence or absence of a disease or condition. Connecting phenotype and genotype data provides information about the genes that may be involved in a disease process or condition, which can be critical for better understanding the disease and for developing new diagnostic methods and treatments [7].

These are powerful examples of products arising from the collaboration of domain scientists with computer and information scientists. With the knowledge and skills needed to use these e-science resources and tools, librarians can support and connect with scientists in their own settings. These products are also brilliant examples of the infrastructure made possible because of NLM. Health sciences librarians can be, in effect, a field force in deploying these

tools and resources to assist the advance of biomedical research.

Health sciences librarian involvement in e-science, then, is essentially made possible by NLM's investments, research, and development. We can be thankful for this leadership. But we should not be complacent and assume that we do not have any responsibility to carry out this work—thinking that NLM has that covered for us—lest we lose touch with the researchers at our institutions and be seen as increasingly peripheral to them. The investment that many health sciences libraries have made to support bioinformatics is one step in the direction of engaging with the research community and providing new services. Learning where the pockets of e-science activity are in your health sciences center—whether in clinical research, basic sciences, or a CTSA program—is a way to continue to span boundaries, understand emerging information management needs, and continue to keep the services we deliver fresh and relevant.

Neil Rambo, nrambo@u.washington.edu, Acting Associate Dean, University Libraries, and Acting Director, Health Sciences Libraries, University of Washington, Box 357155, Seattle WA 98195-7155

References

1. Hey T, Hey J. E-science and its implications for the library community. Libr Hi Tech. 2006;24(4):515–28.
2. Association of Research Libraries, Joint Task Force on Library Support for E-Science. Final report [Internet]. Washington, DC: The Association; 2007 [cited 20 Feb 2009] <http://www.arl.org/rtl/escience/eresource.shtml>.
3. National Science Foundation, Cyberinfrastructure Council. Cyberinfrastructure vision for 21st century discovery [Internet]. The Foundation; Mar 2007 [cited 20 Feb 2009] <http://www.nsf.gov/od/oci/CI_Vision_March07.pdf>.
4. National Science Foundation, Office of Cyberinfrastructure, Directorate for Computer and Information Science and Engineering. Program solicitation: NSF 07-601: sustainable digital data preservation and access network partners (DataNet) [Internet]. The Foundation [cited 20 Feb 2009]. <http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm>.
5. National Institutes of Health, National Center for Research Resources. Clinical research resources: Clinical and Translational Science Awards [Internet]. Washington, DC: The Institutes; 2006 [cited 20 Feb 2009] <http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/>.
6. Arms WY. Cyberscholarship: high performance computing meets digital libraries. J Electron Publ [Internet]. 2008 Winter;11(1) [cited 20 Feb 2009]. <http://hdl.handle.net/2027/spo.3336451.0011.103>.
7. National Library of Medicine. NIH launches dbGaP, a database of genome wide association studies [press release] [Internet]. Bethesda, MD: The Library; 2006 [cited 20 Feb 2009]. <http://www.nlm.nih.gov/news/press_releases/dbgap_launchPR06.html>.